Update from the Data Working Group

Naveed Aziz, Chair, CFI Data Working Group
Vice-President, Research and Innovation, Genome Canada

2026 National Research Facilities Workshop

INNOVATION
Canada Foundation for Innovation
Fondation canadienne pour l'innovation

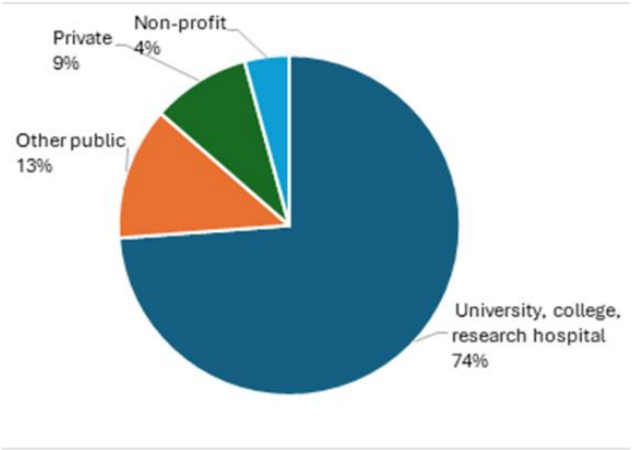
Background

Researchers rely heavily on Canada's 19 national research facilities

- In FY 2024-25 facilities served

17.3 k **on-site and remote users**

2.2 M **data users**



The highest proportion of users (74%) comes from the academic sector

- Since 1997, the CFI has invested nearly

$800 M in capital to build the capacity of these facilities.

❖ Yet 2024 data survey results show **significant gaps** in meeting the data management and governance needs of the facilities

Data Working Group Members

Naveed Aziz (chair)
Vice President, Research & Innovation, Genome Canada

Lam Pho
Chief Information Officer, Canadian Cancer Trials Group

Natalie Harrower
Executive Director, Canadian Research Data Centre Network

Richard Wintle
Associate Director, CGEn Toronto

Tanja Niemann
Executive Director, Consortium Érudit (Coalition Publica)

Sujeevan Ratnasingham
Director, Informatics, Centre for Biodiversity

Benoît Pirenne
Director, User Engagement, Ocean Networks Canada

2025 Data Survey – National Research Facilities

2025 follow up survey conducted to:

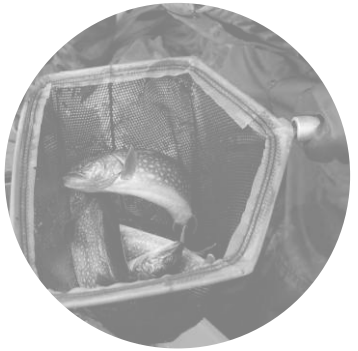
- ✓ **Better quantify the specific data lifecycle needs** of national research facilities over the next two to five years.
- ✓ **Support evidence-based discussions** regarding national digital infrastructure needs with the Alliance and ISED.

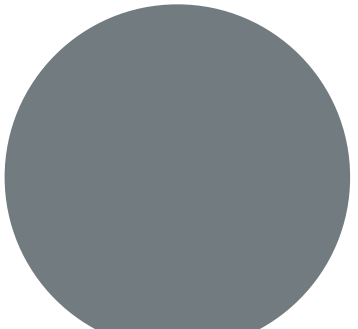
Survey topics

- High performance computing
- Cloud computing
- Data storage
- Software acquisition/development
- Data product development
- Hardware

18 of the 19 national research facilities submitted responses to the survey.








High Level Summary of Needs

High performance computing

Existing facility resources

10 reported CPU cores – average 2,592, range 20-20,000 CPU cores
8 reported GPU cores – average 23, range 7-72 GPU cores
7 reported compute memory (RAM) – average 46 TB, range 1-263 TB

Needs

- Modernized HPC systems to replace aging or obsolete clusters.
- Reduced **wait times** and improved scheduling capacity, especially for large jobs.
- Higher **storage throughput** (I/O)
- Significant expansion of **GPU resources** to support AI and data-intensive workloads.
- Greater access to **scalable shared or cloud-based HPC** to handle peak demand.
- More flexible and responsive **national/shared HPC allocation models**.

Cloud computing

Needs

- More scalable, flexible cloud environments (autoscaling and distributed access)
- Enhanced **data locality, sovereignty, and compliance** solutions.
- More robust cloud-based **disaster recovery** and high availability features.
- Greater reliance on cloud for **AI/ML, analytics, and virtualized environments**.
- **Improved integration** between cloud and on-prem systems for hybrid workflows.

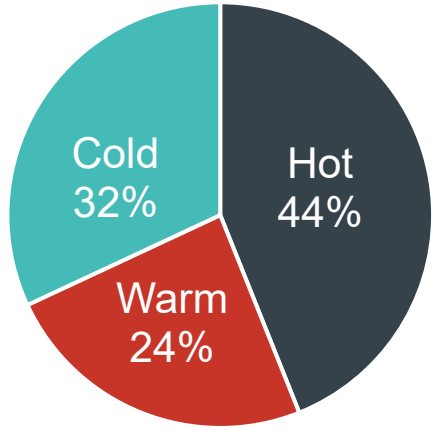
Data Storage

Existing facility resources

Current storage: **~175 PB**

Ranging from **~2 TB** to **~62 PB** at different facilities

Average tier distribution (current)



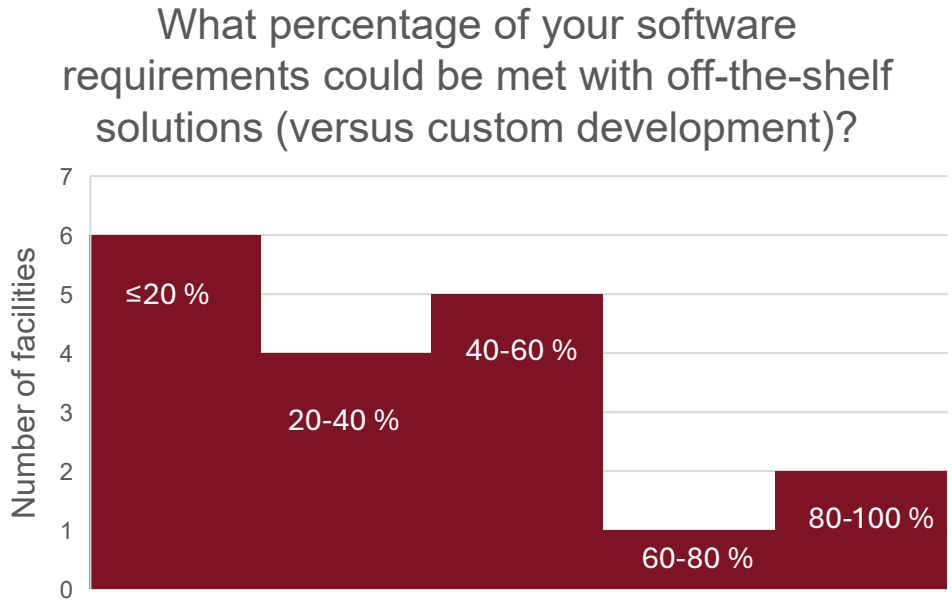
Needs

- Increased storage needed: 2-year**:** **+44.5 PB** | 5-year: **+111.5 PB**
- Need for rapid **multi-PB-scale capacity expansion**.
- More effective **tiered architectures** (hot/warm/cold).
- Higher redundancy and resilience for critical datasets.
- Hybrid storage approaches balancing performance, sovereignty, and archiving.
- Support for **extreme-throughput** workloads (imaging, genomics, AI).

Software acquisition and development

Needs

- Better integration and standardization across tools and workflows.
- **Increased automation and AI-driven metadata/analytics** to reduce manual work.
- Modern, secure software environments (IAM, monitoring, security).
- Continued **custom development** for specialized scientific workflows.
- Enterprise-grade automation/orchestration to reduce technical debt.
- **Shared procurement models** for commercial tools to reduce costs.



*Most facilities can meet **less than half** of their software requirements with off-the-shelf products*

Data Product Development

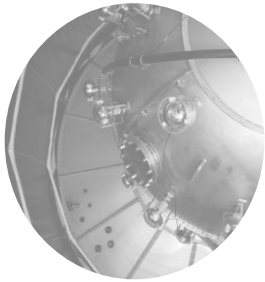
Current output

~3.9 million data products / year produced by facilities
(e.g. dashboards, visualizations, models)

Needs

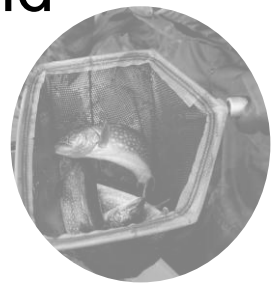
- **Scalable pipelines** for producing, updating, and maintaining large volumes of data products.
- **Automation** in deployment and versioning, especially for frequently updated products.
- Stronger support for **long-term stewardship** of data products with multi-decade lifecycles.
- Systems **enabling high user reach**, both internal and external.

Other Computing Hardware

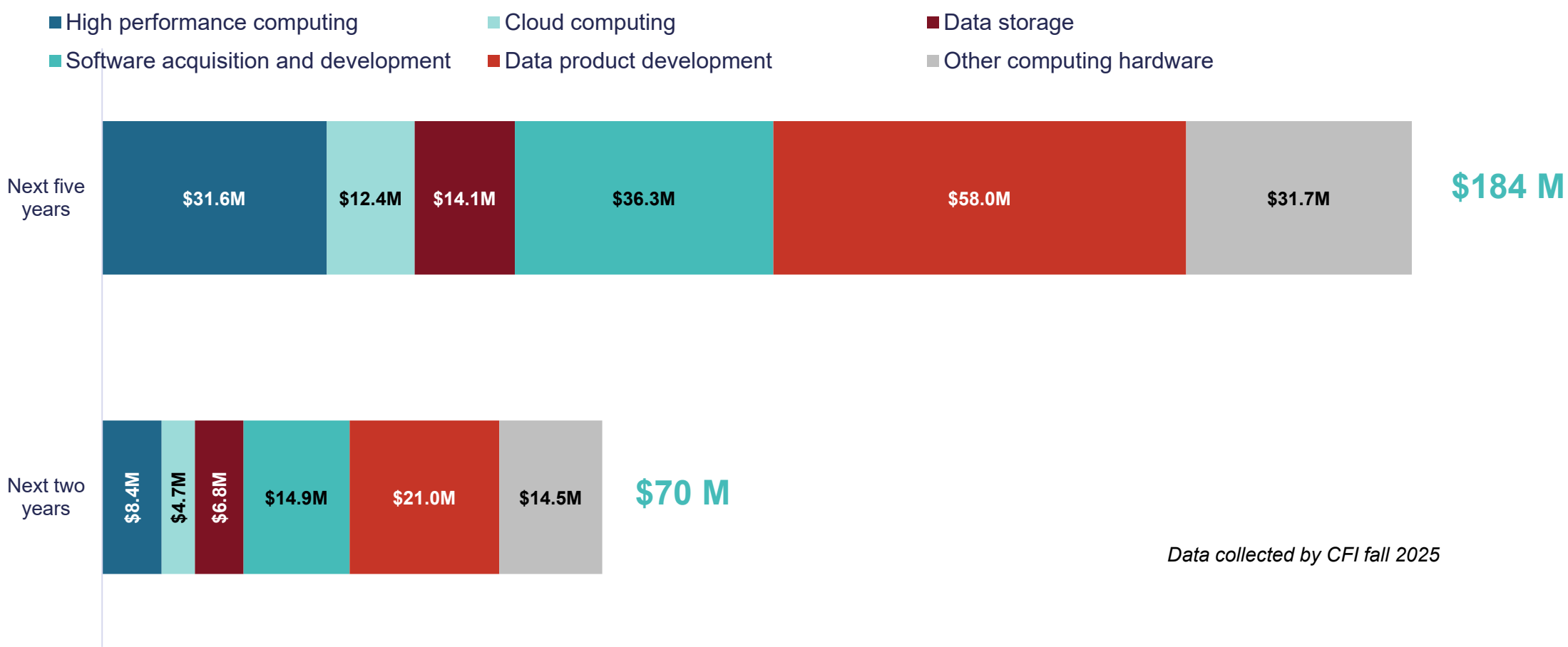


Needs

- Expanded **compute, GPU, and memory capacity** to support modern research workloads.
- **Higher-performance storage and networking** to eliminate I/O and latency bottlenecks.
- Upgraded **power and cooling** to support dense compute environments.
- **Regular refresh cycles** to avoid obsolete or unsupported equipment.
- Secure and **sovereign infrastructure** options for sensitive data.



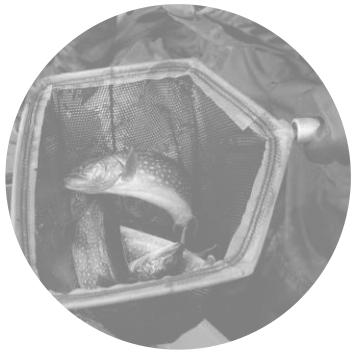
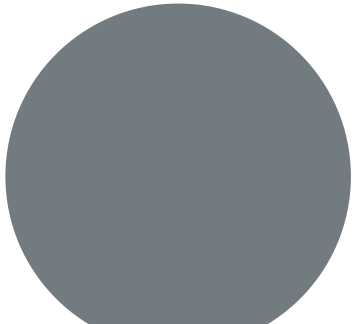
Projected data lifecycle costs for MSI portfolio of national research facilities across the next two to five years



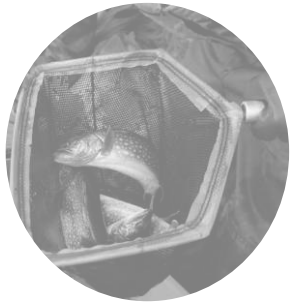
Data collected by CFI fall 2025

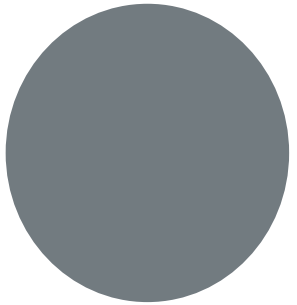
Digital Research Alliance of Canada

- An **essential service provider** for the community.
- Some modernization is required for the Alliance to meet the **unique needs and scale of national research facilities**:
  - **Guaranteed, commercial level** of service for national research facilities, which are not appropriately served by the same mechanisms as individual researchers.
  - More **responsive and flexible allocations**.
  - Better support for **facility-specific workflows**.
  - **Expanded** storage and performance capabilities.

What Does It All Mean?





Three Things You Should Take Away:

1. **Our facilities are world-class, but their data infrastructure needs modernization.** The gap is real, quantified, and solvable. This is urgent.
2. **This is a shared responsibility.** CFI, the Alliance, ISED, and the facilities need to partner on solutions that work at national scale—not piecemeal fixes.
3. **It's fundable and strategic.** A targeted and focused investment over the next five years can modernize data infrastructure for all 19 national facilities. That's the investment in Canada's research future.

Moving Forward

- Input into **Canadian Council of Academies report** on *Enhancing Canada's National Research Infrastructure*
- Continued discussions between CFI, the Data Working Group, and the Alliance to find and implement solutions